Risk Factor Fusion for Predicting Multifactorial Diseases

James Phegley, Kyle Perkins, Lalit Gupta, & J. Kevin Dorsey
Southern Illinois University
Carbondale, IL 62901

Abstract - **A generalized classification methodology which employs risk factor fusion, normalization, DKLT based transformation, feature selection, and parametric classifier design is developed to predict the presence or absence of a multifactorial disease. The validity of method is demonstrated by applying it to predict the occurrence of gout in patients.**

1. INTRODUCTION

The goal in this paper is to develop a data-fusion based pattern recognition approach to predict the presence or absence of a multifactorial disease from a set of risk factors thought to be correlated with the disease. The prediction problem is formulated as a 2-class classification problem in which the two classes are disease and no-disease. The approach developed in this paper includes risk factor fusion, normalization, DKLT based transformation, feature selection, and parametric classifier design. In order to demonstrate the validity of the approach, the prediction of gout, which is a multifactorial disease, is considered. The goal is to classify a patient into one of two classes: gout or non-gout. The approach for gout classification is summarized in Figure 1.

2. DATA FUSION AND NORMALIZATION

The risk factors of a mutifactorial disease can be fused by forming a vector in which each element v_i of the vector is a risk factor. If s is the number of risk factors, let the feature vector consisting of the fusion of the s risk factors be denoted by $V = \{v_1, v_2, ..., v_s\}$. Each patient, therefore, is represented by a single feature vector. Let the feature vector of the patients with the disease be denoted by $V^m = \{x_1, x_2, ..., x_s\}$ and those who do not have the disease be $V^n = \{y_1, y_2, ..., y_s\}$, respectively. Additionally, let the feature vector of a test patient be $V_t = \{t_1, t_2, ..., t_s\}$. In general, the feature vector will contain features with mixed formats because the risk factors can be real-valued with widely differing ranges as well as binary factors. In order to facilitate classifier development, the features can be normalized. For example, real-valued features can be linearly normalized to take real values in the interval [0.1, 0.9] and for the binary features, zeros and ones can be assigned values 0.1 and 0.9, respectively. The motivation for developing this normalization approach is to not only accommodate features of mixed formats but also to initially assign equal weightage to all features. That is, although it is known that the risk factors are likely to have varying degrees of correlation with the disease, no assumptions are made initially about the influence of these factors on the prediction of the disease.

3. CLASSIFIER DEVELOPMENT

Deciding whether a test patient has or does not have the disease can be formulated as a hypothesis testing problem in which:

$H_0 : V_t = V^n$; the patient does not have the disease.

$H_1 : V_t = V^m$; the patient has the disease.

In order to facilitate classifier development, the feature vector which has both real and binary features, can be transformed so that the transformed feature vector contains real-valued features. As a result, it would be much easier to make meaningful assumptions for the conditional densities of the feature vectors under the two hypotheses. The discrete Karhunen-Loeve transform (DKLT) is a suitable transformation because each feature in the transformed vector is a linear combination of the features in the original feature vector. Let $\tilde{V}^n$ and $\tilde{V}^m$ be the transformed features vectors. That is,

$$\tilde{V}^n = \Phi V^n$$
$$\tilde{V}^m = \Phi V^m$$

where, Φ is the generalized DKLT transformation matrix consisting of the eigenvectors of the covariance matrix of the mixture of the 2 classes [1]. The transformed features, which are weighted combinations of the original features, can be rank ordered in terms of the inter-class separation in order to select the features which are the most useful for separating the two classes. If $\tilde{x}_{ij}$ and $\tilde{y}_{ij}$ are the *ith* components of the *jth* transformed feature vector for the disease and no-disease classes, respectively, then, the inter-class separation between the trans-

0-7803-7211-5/01$10.00©2001 IEEE

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Risk Factor Fusion for Predicting Multifactorial Diseases	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Southern Illinois University Carbondale, IL 62901	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-26, 2001 held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

formed features $\tilde{x}_i$ and $\tilde{y}_i$, $i = 1,2,...,s$, is given by

$$\beta[i] = \frac{[\bar{\tilde{x}}_i - \bar{\tilde{y}}_i]^2}{\{\sum_{j=1}^{J}[\tilde{x}_{ij} - \bar{\tilde{x}}_i]^2 + \sum_{j=1}^{J}[\tilde{y}_{ij} - \bar{\tilde{y}}_i]^2\}},$$

$$i = 1,2,...,s,$$

where, $\bar{\tilde{x}}_i$ and $\bar{\tilde{y}}_i$ are the mean values of x_i and y_i, respectively, and J is the number of feature vectors in each class.

Let $\hat{V}^n$ and $\hat{V}^m$ be the truncated feature vectors formed by selecting the k components with the highest separations between the features in $\tilde{V}^n$ and $\tilde{V}^m$, respectively. The two hypotheses can, therefore, be written as

$H_0: \hat{V}_t = \hat{V}^n$; the patient does not have the disease.

$H_1: \hat{V}_t = \hat{V}^m$; the patient has the disease

The truncated test feature vector $\hat{V}_t$ is given by

$$\hat{V}_t = \hat{\Phi} V_t,$$

where, $\hat{\Phi}$ is the matrix formed by retaining the k eigenvectors in the matrix Φ that yield the k highest separation features. If the a priori probabilities $p(H_0)$ and $p(H_1)$ are assumed equal, the likelihood ratio decision rule can be written as,

$$\Lambda = \frac{p(\hat{V}_t / H_1)}{p(\hat{V}_t / H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{p(H_1)}{p(H_0)} = 1.$$

At this point, assumptions can be made for the conditional densities $p(\hat{V}_t / H_0)$ and $p(\hat{V}_t / H_1)$. For example, if it is assumed that the conditional densities are Gaussian and if $\mathbf{M}_n$ and $\mathbf{M}_m$ are the mean vectors and Ψ_n and Ψ_m are the covariance matrices of $p(\hat{V}_t / H_0)$ and $p(\hat{V}_t / H_1)$, respectively, the likelihood ratio decision rule after taking logarithms and rearranging terms can be written as

$$(1/2)[(\hat{V}_t - \mathbf{M}_n)^T \Psi_n^{-1} (\hat{V}_t - \mathbf{M}_n) - (\hat{V}_t - \mathbf{M}_m)^T \Psi_m^{-1} (\hat{V}_t - \mathbf{M}_m)] \underset{H_0}{\overset{H_1}{\gtrless}} \ln(|\Psi_m| / |\Psi_n|)^{1/2},$$

where, $|\mathrm{A}|$ is the determinant of matrix A.

4. PERFORMANCE EVALUATION

The question of how best to partition a data set into a design set for classifier development and a test set for testing the classifier has received considerable attention [2-4]. The cross-validation method which randomly partitions the data set into two mutually exclusive and equi-sized sets to generate the design set and test set can be employed to evaluate the performance. This method is effective only when the design set is large enough to robustly estimate the parameters of the classifier. The classification accuracy is estimated as the fraction of the number of correctly classified vectors in the test set. The random partitioning can be repeated H times (trials) and the classification accuracy is then estimated by averaging the resulting H classification accuracies. That is, the classification accuracy is given by

$$\alpha_H = [(1/H)\sum_{h=1}^{H} \alpha_h] \times 100\%,$$

where,

$$\alpha_h = \frac{number\ of\ correctly\ classified\ vectors}{number\ of\ vectors\ in\ the\ test\ set}$$

in trial h, h=1,2,…,H.

5. GOUT

Gout is a form of arthritis usually caused by increased levels of uric acid circulating in the blood and being deposited as needle-like crystals within joints and tissues. These deposits lead to episodes of inflammatory arthritis which results in pain, swelling, redness, and damage to the joints. Elevated uric acid levels alone, however, are not sufficient to diagnose gout because only 10% to 20% of individuals with high levels of uric acid develop gout [5]. Additionally, the uric acid levels in the blood may be transiently normal or low during a gout attack [6]. Gout may be difficult to diagnose at times because symptoms may mimic other rheumatic diseases [7]. Conversely, other kinds of arthritis can mimic a gout attack. Because treatment of gout is specific, the correct diagnosis is essential.

Risk factors for gout have been studied extensively for years and the correlations between the risk factors and gout are summarized in reference [8]. The following 14 variables from the risk factor set were included in this study: serum uric acid, gender, age (at diagnosis of gout), the presence or absence of diabetes, the presence or absence of hypertension, weight, height, body surface area, history of kidney stones, the presence or absence of thiazide diu-

retics, serum cholesterol, triglycerides, creatinine, and blood urea nitrogen.

6. GOUT CLASSIFICATION EXPERIMENTS

The computer records of a multi-specialty group practice were searched for patients with a diagnosis code for gout who had an office visit during a nine-month period. Of 91 charts available for review, 48 patients were identified who had information available for all parameters under investigation. The diagnosis of gout was considered confirmed by a rheumatologist if either uric acid crystals were identified in synovial fluid, or a classic attack of podagra (acute arthritis of the great toe) was documented, or if there were typical recurrent attacks of monoarticular arthritis in the absence of another arthropathy which could account for the symptoms.

Forty-eight patients without gout were matched for gender and for age to the gout patients and were identified from the clinic laboratory records as having had a multi-channel chemistry profile performed during a previous three-month period. In both cohorts, patients were excluded if they had psoriasis, or a lymphoproliferative disorder, or were taking uricosuric therapy or allopurinol at the time the laboratory data were available because all could alter their serum uric acid values. The control group excluded those patients taking aspirin (lowers serum uric acid), but patients taking other non-steroidal anti-inflammatory drugs were included.

Classifiers to predict the occurrence of gout were developed exactly as outlined in Sections 2 and 3. The cross-validation method described in Section 4 was used to evalute the performance. The estimated classification accuracies of the Gaussian classifiers, expressed as percentages, are shown in Figure 2. The results are presented for H=50 and for k=1,2,…,14, where, k is the number of rank ordered features in the transformed feature vectors. For each of the H=50 random partitions into a training and test set, the DKLT transformation matrix was computed from the training set and the test results are shown by the solid lines in Figure 2. The results show that an average classification accuracy of 75.7% can be obtained by selecting the first three highest inter-class separation components in the transformed feature vectors. The dotted lines in Figure 2 show the test results obtained by selecting, through trial and error, a DKLT transformation matrix that gave good results. That is, the DKLT transformation matrix was computed just once and tested H=50 times with randomly selected test sets. It is seen that by selecting the first 5 highest inter-class separation components, a classification accuracy of 87.4% can be achieved.

7. CONCLUSIONS

This paper focused on developing a pattern recognition methodology to detect the presence or absence of a disease from a set of risk factors correlated with the disease. The generalized classification methodology developed included fusion to combine risk factors into a single feature vector, normalization to overcome the problems associated with fusing features which have different formats and ranges, DKLT based transformation to facilitate parametric classifier development, feature selection, and Gaussian likelihood ratio classifier design. The methodology was applied to detect the presence or absence of gout from a set of 14 risk factors thought to be correlated with gout. Cross-validation evaluations on patients clinically diagnosed to have gout and not have gout showed that on the average, a classification accuracy of 75.7% can be obtained which is quite encouraging. What is even more promising is that classification accuracies of over 87% can be achieved through the careful selection of the DKLT transformation matrix which in turn involves selecting training sets that are good representatives of the gout and non-gout classes. It could, therefore, be concluded that the gout and non-gout classes could be separated even more accurately if larger and more representative training sets were available for the two classes. In summary, the generalized disease prediction methodology developed in this paper can assist a physician in diagnosing a multifactorial disease.

REFERENCES

1. J. T. Tou and R.C. Gonzalez, *Pattern Recognition Principles*. Addison-Wesley, Reading, MA, 1974.
2. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, Chapter 3, 1973.
3. W.J. Krzanowski, *Principles of Multivariate Analysis*, Oxford Science Publications, Chapter 12, 1988.
4. K. Fukunaga, *Introduction to statistical pattern recognition*, Academic Press, 1990.
5. F. Wolfe, “Practical therapeutics – gout and hyperuricemia,” American Family Physician, 43, 2141-2150, 1991.
6. R.J. Glynn, E.W. Campion, & J.E. Silbert, “Trends in serum uric acid levels

1961-1980," Arthritis and Rheumatism, 26, 87-93, 1983.

7. F. Wolfe, & M.A. Cathey, "The misdiagnosis of gout and hyperuricemia," Journal of Rheumatology, 18, 1232-1234, 1991.

8. K. Perkins, B.D. Wright, and J.K. Dorsey, "Using Rasch measurement with medical data," in Rasch Measurement in the Health Sciences, Nikolaus Bezruczko, Ed, Chicago: MESA Press, University of Chicago, in press.

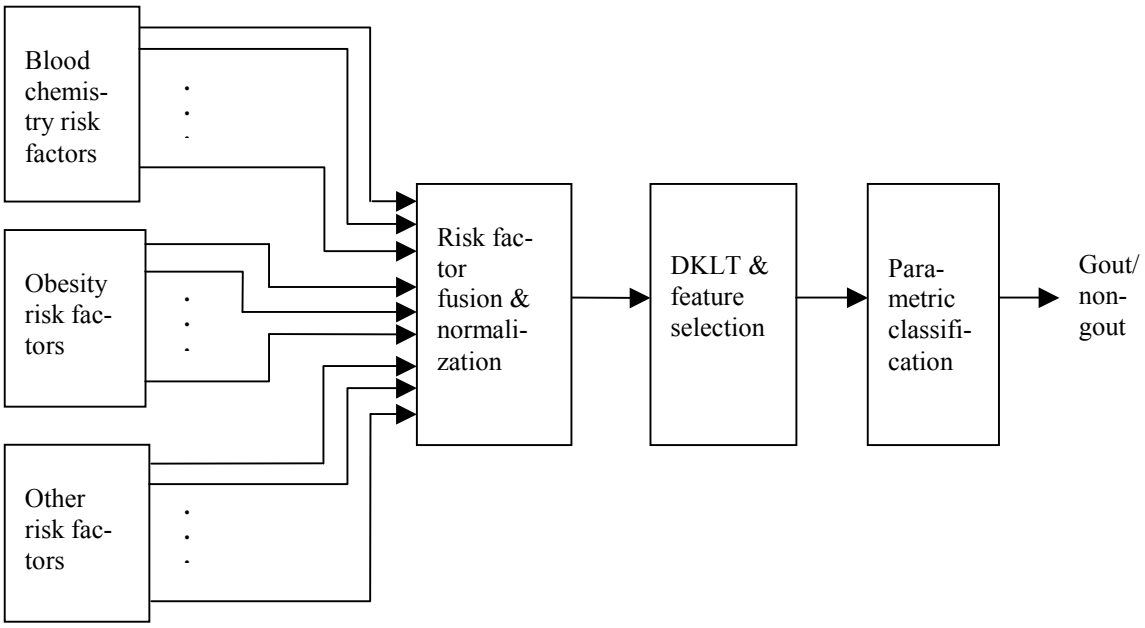

Fig. 1. Block diagram of the risk factor fusion approach to classify gout.

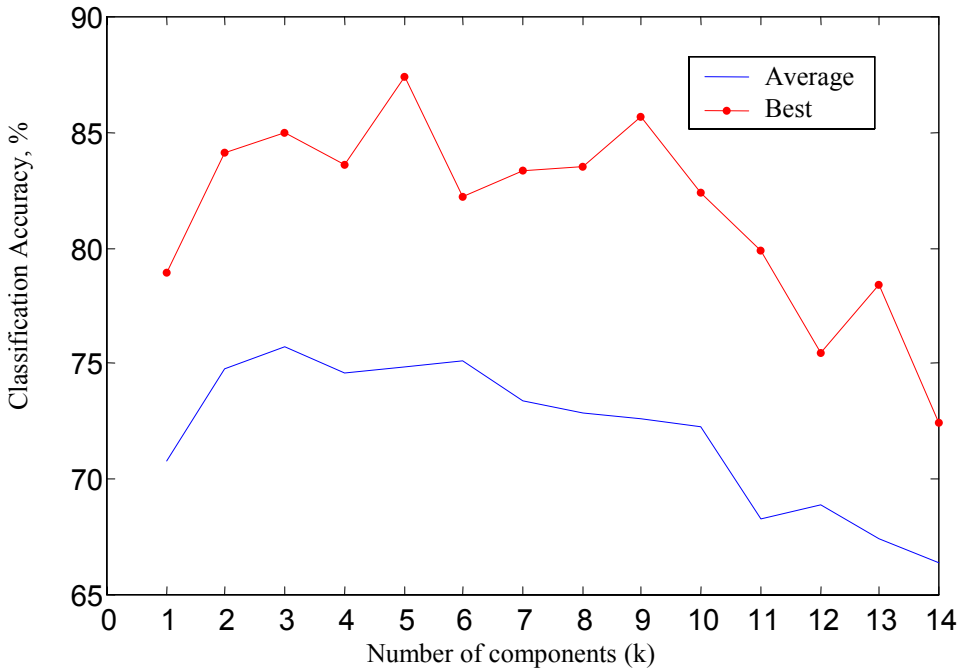

Fig. 2. Classification results.